19 May 2024

Natural Language Processing on the Official Gazette of the Republic of Turkey for Detection of Agricultural and Rural Topics for Understanding Food Security Framework

Orhan Akpınar - CSSM530 - Automated Text Processing for Social Sciences
Koç University - Computational Social Sciences MA

Course Instructor: Dr. Ali Hürriyetoğlu, Koç University (CSS)
Domain Expert: Dr. Cemil Yıldızcan, Galatasaray University (Political Sciences)

Resume

Abstract: Food security is a problem that has multifaceted causes and effects from the ecological to the public sphere. Legislations covering agrifood systems are important for stakeholders' and consumers' sustainability of well-being. However, legal documents are hard to organize, interpret, and measure for those who are outside of expert communities. A more democratic approach would be incorporating multiple levels of stakeholders, as well as consumers, for being an active participant in the legislative framework of agrifood systems. In this paper, I will design a topic classification model for the Official Gazette of the Republic of Turkey. This study later aims to help both non-data science experts and food security stakeholders by facilitating information retrieval and analysis regarding newly imposed laws as well as comparing previous laws' consequences.
Method: Natural Language Processing for Topic Detection using Annotated Data
Data: Official Gazette of the Republic of Turkey from 26 July 2006 to 18 May 2024
Results: 97% F1-macro on the consecutive years, using 2012 to train and 2013 to test the predictions. 87% F1-macro score for using these two years to predict 3 over 47 randomly selected titles between 26 July 2006 and 18 May 2024 excluding training years.
Conclusions: Topic detection can be useful for the empirical analysis of the legislative framework, it would facilitate data acquisition and increase collaboration among expert researchers.
Contributions: Stored the content of the Official Gazette of the Republic of Turkey, and published openly more than 1600 annotated entries. The work is replicable and improvable, and may provide a cross-domain approach to the socio-political analysis of legal texts.

Note: Post-Review Notes are added at the beginning of this paper.

Post Review Notes

Annotation:

As stated in the paper, resource constraints require me to be the only annotator. The annotation model has been mentioned in the annotation manual.

Error Analysis:

There are two cases related to errors. One is due to model inaccuracy due to lack of training data. As mentioned in the annotation manual, the training dataset did not contain future relevant cases. Therefore, new predictions were giving false negatives. On the side of false positives, the appearance of some words with target phrases caused mistreatment of the newer appearance. The validation requires handling legal terminology. Therefore it is true that scaling the process will be hard to tackle. The suggested way can be training on random data containing multiple years, however, even this process cannot ensure the variation. Secondly, there might be errors due to the machine learning algorithm. Preprocessing with TF-IDF and applying different hyperparameters might increase the accuracy of the model.

Limited Scope:

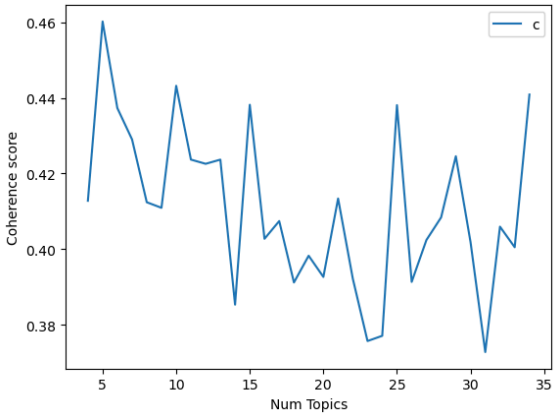
As discussed in the data chapter, computing resources prevent acquiring the full scope of the laws. However, on the technical side, this project has been done with CPU resources. In the next step, using a CUDA-compatible GPU will strengthen the scope of the study.

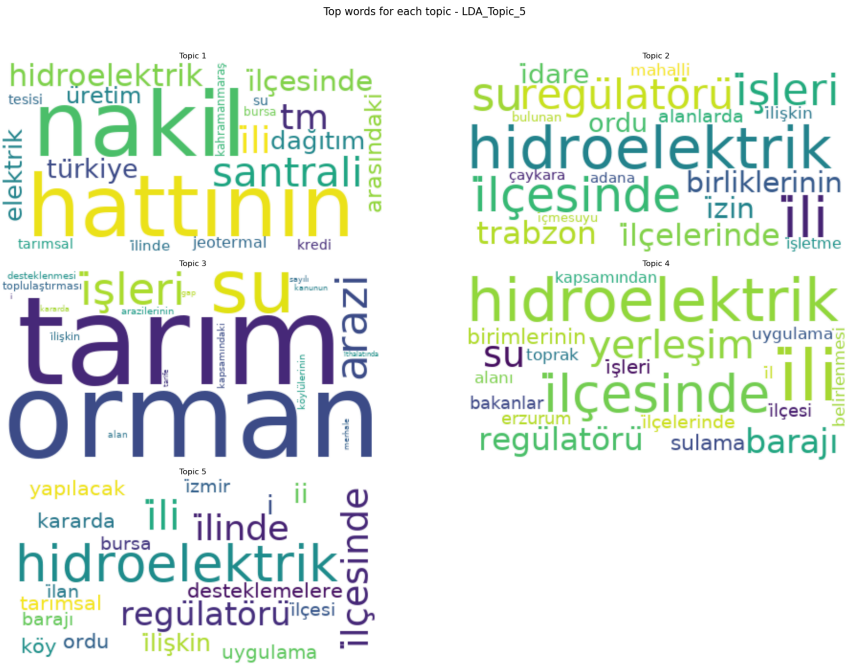
Notes From Domain Expert:

We have discussed the case that,

1. Capturing obscure laws can be very useful. As mentioned in the annotation manual such laws stating “Law About Defining the Application Dates of Some Laws” (Bazı Kanunların Uygulanma Tarihlerinin Belirlenmesi Hakkında Kanun) type laws are incomprehensible. Detecting the content of such laws can be very beneficial.

2. As the later suggestions provided by Dr. Cemil Yıldızcan, I eliminated laws stating communization (“kamulaştırılması”) from the second version. In this manner, I tried to simplify the topic modeling by focusing only on laws about water, soil, agriculture, and rural policies.
3. By the newly annotated set, I applied the same procedure. I validated 2012, then predicted 2013, then validated it. In the last case, F1-macro score dropped to 75%. Considerations that, since I eliminated more target data, the distribution became more skewed. This might cause the inaccuracy. However, later investigation is required.
4. Finally, LDA results are hereby presented regarding last results:





Introduction

Food security has been defined with over 200 different definitions (Maxwell 1992). In simple words, the term regards the sustainability and accessibility of healthy and sufficient food. For this reason, agricultural production and rural livelihood present an important dimension of food security. Government subsidies and rural governance laws pose a variation of benefits and concerns for the producers and consumers. More importantly, imposed laws cannot be easily predicted for their intended effects. The reason is that evaluation of law effects touches upon social, economic, and ecological issues. As a result, measuring the outcome of laws requires a collaborative approach from multiple domains of expertise.

On the other side, outside of the expert community, consumers and producers are active participants in agrifood systems. However, a lack of expert knowledge prevents a careful understanding and participation in governance. As a controversial example, government subsidies for supporting agricultural producers do not directly increase the agricultural output but rather increase the possibility of taking higher risks in crop selection and maintenance, however it is still positively affects. (Günaydın 2009). This particular issue then creates higher chances of loss in agricultural production. Another example would be subsidies for hybrid seeds, which reduce domestic seed production and create a dependent agricultural sector (Lee 2007).

A plausible explanation may be given regarding the disconnection between monetary input, knowledge input, and domestic organization of crop type selection. Therefore, analyzing a single framework of law without carefully defining other legislative frameworks regarding agriculture and rural livelihood will create an omitted variable problem for analyzing the effects.

This problem affects especially small-scale farmers, who are the most fragile against adaptation to new legislative frameworks, ecological conditions, and agricultural market dynamics (Lal 2023). Similarly, retail consumers are as distant to policy frameworks and their effects as small-scale farmers. The term food sovereignty underlines an additional topic regarding non-corporate participants in the agrifood systems (Lee 2007).

Data

I will use the Official Gazette of the Republic of Turkey[1] to capture the laws of interest because all new law publishings as well as other types of official announcements released daily except on holidays. Data is accessible from the website, covering all 32545 issues from 7th February 1921.

Each issue includes multiple titles, and titles can be syntactically plural when there are multiple contents under them. Such that “Kanun” (Law) and “Kanunlar” (Laws) refer to the same category but with different numbers of contents. Besides, some titles are without content, such as “Yönetmelikler”**, and all content is included under the same title. In my case, I will be interested in the categories – "Kanun(lar)", "Bakanlık Kurulu Karar(lar)ı" and "Cumhurbaşkanı Karar(lar)ı" due to their legal content. Other categories are recorded in scraped_data files but excluded from filtered_data files which are used for training and validation.

Noting that this work will not provide an in-depth analysis of the published announcements, but rather will help to facilitate detection and compilation of laws for an in-depth analysis. In my focus, on researching the food security aspects of legislative frameworks, I will train the NLP model for detecting agricultural and rural legislation. From a preview of publishings of

[1] https://www.resmigazete.gov.tr/

the year 2012, I saw that such laws can be related to production, trade, rural management, and natural resources.

The reason that I am not proceeding to perform an in-depth analysis of published laws is due to the extensive resources required to process the full issues. From a sample of 10 days a compiled PDF document can be 1700 pages long. Additionally, not all pages are text-processable and require an optical character recognition pipeline for processing. Consequently, the lack of computational resources for the extensiveness of this kind of research led me to leave it for future research.

Therefore, I decided to take a step back to make the research easily replicable and adjustable. Although the content is lost, such as the law with the title *"On Üç İlde Büyükşehir Belediyesi ve Yirmi Altı İlçe Kurulması ile Bazı Kanun ve Kanun Hükmünde Kararnamelerde Değişiklik Yapılmasına Dair Kanun"* does not define which thirteen cities are subject to the law. However, an elementary level of domain knowledge will be enough to skim the article for understanding the topic. If needed, the use of summarization models can further speed up the process for larger in-depth document processing. This topics are considered in discussion and future research.

Data Cleaning

For acquiring and cleaning the data, with the categories mentioned above, I used web scraping and html parsing with BeautifulSoup4[2] to access the categories and subtitles. The first filtered acquisition test complied with 576 subtitles from 117 days. I proceeded with the acquired subtitles for annotation and training, followed by the full acquisition of years 2013 for initial testing and validation. After the initial train-test-validation process, I proceeded with issues from 20 Feb 2006 to 17 May 2024. The year selection of test steps is chosen due

[2] https://beautiful-soup-4.readthedocs.io/en/latest/

to the scraping model. An additional testing method can include random dates between 1921 and 2024 for general applicability tests. However, different dates will require different parsing methods.

Data Annotation

The annotations are conducted by myself, therefore annotation methodology requires further validation. An annotation manual requires experts from multiple domains because food security issues are multifaceted. Such that, it might be hard to understand a law about water irrigation by an economist, or a municipal governance law by an ecologist, et cetera. For example, municipal organizations might be unrelated to food security but also can be related to urban farming practices. Similarly, a natural disaster news in an agricultural city is not directly related to food security but can be related to it due to its consequences for producers. Therefore, I decided to incorporate everything that contains a framework related to rural or agricultural organizations. Even so, I might fall into over-specification of my dependent variable and acquire redundant returns. As a result, this work will not be able to provide a relational analysis of such laws but is aiming to create a collaborative and robust framework for in-depth analysis.

Selected Phrases

The following list contains phrases extracted from the 2012 filtered dataset while conducting the annotation. The dataset contains 576 entries which are annotated manually.

Kamulaştırılma(all content included); Köy; Orman; Tarım; Su; Ziraat; Toprak(Normu); Baraj; GAP; Güneydoğu Anadolu Projesi (appeared with “Damızlık Sığır” in2013, although it is related to livestock, it is also included because it is mentioning the project.); (Bozulabilir) Gıda; Arazi Toplulaştırılması; Şeker Kurulu; Kota (of food production, coexist with sugar); MahalliİdareFalse positives are possible for bodies related to tourism, trade etc.; false

negatives seen in sample1, sample2); Patates; Çiftçi; Kuraklık; Büyükşehir Belediyesi; Rize Serbest Bölgesi (due to its relation to Tea production)

The following list contains phrases that are seen in the validation of the 2013 filtered dataset. Some phrases indicate false negatives, some phrases are recorded due its new combination of phrases. Please refer to validation and false_pred datasets for checking details.

Bitki; Tarım Ürünlerinde Taviz; Çevre Alanında İşbirliği; Tarımsal İşbirliği;Tarım Alanında İşbirliği; Fındık ; Mera; Tarım ve Hayvancılık Bakanlığı (which actually contains information related to veterinary issues, recorded due to the named entity);Uygulama Alanı(anunexpected true positive. Never seen in the 2012 annotation, but it is related to “Kamulaştırma” or “Köy” due to land use content. In the first annotation set from 2012, “Uygulama Alanı” is annotated as 1 because it coappears with “Toprak” or “Köy”. In the example of 2013, it appears with “Bazı Yerleşm Birimlerinin”, with no attribution to other topics; also got a false negative in 2008, sample17); Atatürk Orman Çiftliği

Finally, the below list shows false negatives that are seen in randomized validation sets. The original set contains data from 27 July to 17 May 2024, omitting 2012 and 2013 issues. The complete filtered dataset has 5029 entries. This dataset was randomly sampled in 47 datasets, each containing 107 entries. I validated samples numbered 1, 2, 17, 41. Files can be seen in the samples file.

Tarım Kredi (2011, sample2; 2011, sample17); Sebze ve Meyve (2010, sample1); Biyogüvenlik(2010, sample1); Doğal Sit Alanı (2020, sample1; 2022, sample2);Hububat(2006, sample17); Özel Çevre Koruma Bölgesi(2006, sample41)

Omitted Phrases

Additionally, to selection rules, I also sketched some of the ommittance choices. This part is required to prevent confusion. Most of the contents which do not contain focus words are generally easy to annotate negatively. However, in some subtitles, the word choice may present misunderstanding. Additionally, this step may be required to deal with a more focused target set.

Tapu; Kadastro; Kentsel Dönüşüm (no “Kamulaştırma” exists); İmar; Ticaret; Milli Savunma; EtilenGlikol; Alkol; Tütün; Taşınmazın[Kazakistan Cumhuriyetine] Kullandırılması (2013); Taşınmazın Bedelsiz Olarak [“Beyoğlu Belediyesine”, a governance institution is stated] Devredilmesi (2013); Veteriner(2013); Kalkınma(generally relevant to industry); Kalkınma Ajansları (2022, sample2); Amortisman; Turizm; Yenileme; Kültür; Ekonomi; Güvenlik; Enerji(included if “Kamulaştırma” exists); Ulaştırma; Uluslararası (included if “Tarım” or “Bitki” or agricultural topic stated); Gümrük (included if agricultural topic stated); Yenileme Alanı (2013); Ticaret Anlaşması; Finansman; Çevre ve Şehircilik Bakanlığı; Pamuk; Özelleştirme(which means selling an asset, not directly indicating agricultural or rural topic); İthalat(If not containing agricultural contentin the title); Yap-İşlet-Devret(generally is a process after “Kamulaştırma” of a land); Bütçe Kanunu; Petrol; Deniz; Riskli Alan İlan Edilmesi (2013); Kültür ve Tabiat Varlıklarını Koruma Kanunu (2013); Jeotermal Kaynaklar ve Doğal Mineralli Sular Kanunu (2007, sample17)

Lastly, the below list indicates the obfuscated subtitles. It is not possible to understand what they refer to just by looking at the titles. Therefore they are omitted. This part necessitates an additional summarization pipeline.

Bazı Kanun ve Kanun Hükmünde Kararnamelerde Değişiklik Yapılmasına Dair Kanun; Bazı Anlaşmaların Yürürlüğe Girdiği Tarihlerin Tespit Edilmesi; Bazı Kanun ve Kanun Hükmünde Kararnamelerde Değişiklik Yapılmasına Dair Kanun; Kanun Hükmünde

Kararname ile Bazı Kanun ve Kanun Hükmünde Kararnamelerde Değişiklik Yapılmasına Dair Kanun

These are the detected and omitted phrases for defining the target results. Many false cases are not defined here, such related to “Military”, “United Nations” etc, because the focus was on confusable phrases. One important aspect, Büyükşehir Kanunu (Metropolis Law) has a different word choice than other words related to agriculture and rural legislation. Which was the main topic of rural organizations on a great scale comprising thirteen additional cities accepted in 2012.

Training the Natural Language Processing Model

I am using “dbmdz/bert-base-turkish-uncased”[3] as the Turkish BERT model. There are multilingual models as well, which can be tested. Using the trained model on the year 2012 annotated dataset, I predicted and later manually validated the 2013 dataset.

The ten-fold cross-validation prediction score of 2012 is %99.70. I used the annotated 2012 dataset to train the prediction algorithm using BERT, which predicted the 2013 filtered dataset. In the last part, I used validated 2012 and 2013 datasets to train a model for predicting the Official Gazette of the Republic of Turkey issues from (26 July) 2006 until 18 May 2024, omitting 2012 and 2013 to prevent data leakage. 26 July is the date when the structure of the online official gazette changed. Therefore it is suited for my scraping code. At this stage, it is possible to use it for current issues.

[3] https://huggingface.co/dbmdz/bert-base-turkish-uncased

Results and Discussion

There are 576 entries in the 2012 dataset and 691 entries in the 2013 dataset. As a result, the BERT prediction algorithm performed at 0% False Positive, and 4.96% False Negative corresponding to a 0.9746 F1 macro score.

Lastly, using the scraped data from 26 July 2006 to 18 May 2024, I created 47 random samples from 5029 total entries. For each sample containing 107 rows, samples numbered 1, 2, 17, and 41 were validated manually. As a result, I received 0.6% false positives, and 22.61% false negatives corresponding to an F1-macro score of 0.8641. Most of the false negatives occurred unexpectedly while containing phrases “Uygulama Alanı”, “Mahalli İdare”, and “Hububat” the prediction model did not include those. Additionally, a new phrase “Doğal Sit Alanı” and “Biyogüvenlik” were detected while checking false negatives. Although the last term is not related to agrifood systems, I choose to include them in my target set due to their possible relation to rural livelihood and ecosystem health. Newly annotated data has not been tested on the rest of the samples.

Finally, using annotated 2012 and 2013 filtered datasets, I applied a topic modeling algorithm using LDA (Blei 2003) after preprocessing it with spacy model Turkish NLP Suite (Altınok 2023). While checking coherence metrics, clusters from 4 to 35 are tested to achieve the best results. However, there were different results each time, but generally, it is best to use a cluster number between 9 and 14. However, it was also shown to get good coherence results with numbers around 30. The problem is related to defining stopwords in text preprocessing.

The legislative texts contain many common words, and most of them are employed to annotate the titles manually. However, these words posed a problem while applying LDA due to their high frequency. Based on most_weighted_words, I provided stopwords manually in

preprocessing. Any changes in the most used words, however, caused a lot of disruptions in the output. Additionally, while tokenizing the texts, I saw that some words are concatenated during the scraping process. This created additional problems. For future directions, I aim to apply TF-IDF (Robertson 2004) in preprocessing to statistically reduce the noise.

Below, I provide an output for LDA modeling with 12 topics.

Figure.1 : Wordclouds of LDA Topics with 12 clusters

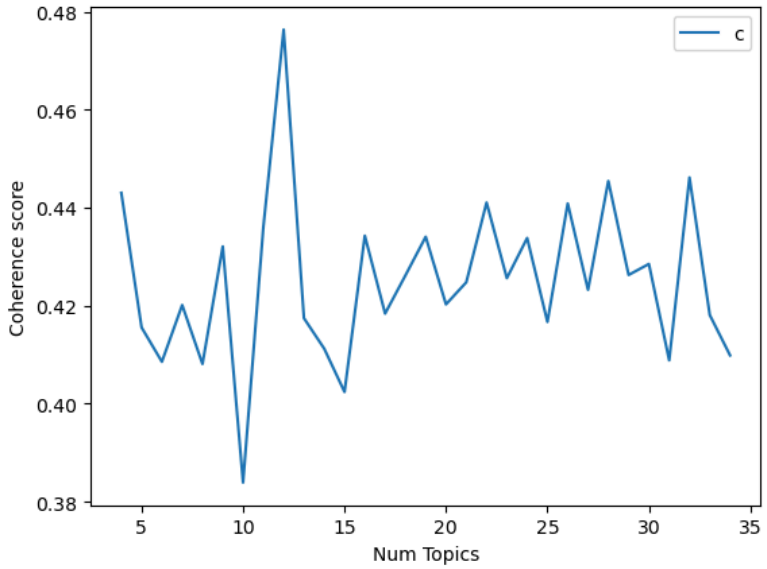


Figure.2: Coherence validation of topic clusters using LDA

Even though the output is not clean, there is a partially retrieved infromation. As indicated in the annotation manual, these words represent different topics for further research. Some examples are: Arazi toplulaştırılması, tarım, orman ("orman köylüsü"), su (along with "baraj", "sulama"), mutabakat (işbirliği), elektrik (which is related to "hidroelektrik santral", and "şirket", which is a major part of "Kamulaştırılma" legislations). There were additional problems in processing Turkish text, which caused malfunctioning of stop-words, causing them to inconsistently apply to texts. For this reason, character encoding to the English alphabet is suggested.

In conclusion, I presented a method to acquire and analyze the Official Gazette of the Turkish Republic, which contains every legal document that is published. This type of service is rather expensive[4], and social scientists from political, social, and economic domains cannot easily search through the contents. Acquiring and thoroughly analyzing the official gazette would be very helpful in understanding the domestic social context.

[4]https://www.lexpera.com.tr/uyelik-destek/paketler-ucretler

Contributions and Future Directions

Hereby, I present the complete scraped datasets from 26 July 2006 to 18 May 2024, which nearly contain 70 thousand entries. From this, there were nearly 6 thousand entries related to my target categories. Among them, I provide more than 1600 validated entries for training a text classification algorithm. The initial training started with 576 entries from 2012, to evaluate 691 entries from 2013. From this process, I got ~97% F1-macro score. On 3 (107 entries each) of 47 randomly sampled datasets from the rest of the target dates, the prediction algorithm returned ~86% F1-macro score cumulatively.

The next target would be applying a summarization model on the target topics. The problem is, that each subtitle contains the text in the link, and each text may contain additional links in the main body for referring to the general body of decisions. Moreover, some content is not parsable directly via text processing and requires optical character recognition. Therefore, this summarization task would require a multiple step of content acquisition, parsing, and cleaning. However, while open-source resources are increasing at a rapid pace, this collaborative task is a must for a deeper understanding of the laws and their social outcomes.

References

1. Altinok, D. (2023, July). A Diverse Set of Freely Available Linguistic Resources for Turkish. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 13739–13750). doi:10.18653/v1/2023.acl-long.768

2. Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. J. Mach. Learn. Res., 3, 993--1022. doi: http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993

3. Günaydın, G. (2009). Türkiye Tarım Politikalarında 'Yapısal Uyum': 2000'li Yıllar. Mülkiye Dergisi, 33(262), 175-221.

4. Lal P., Chandel BS, Tiwari RK, El-Sheikh MA, Mansoor S, Kumar A, Singh G, Lal MK, & Kumar R. (2023). Effects of agricultural subsidies on farm household decisions: a separable household model approach. Front. Sustain. Food Syst., 7, 1295704. doi: 10.3389/fsufs.2023.1295704

5. Lee, R. A. (2007). Food security and food sovereignty. Retrieved from https://www.ncl.ac.uk/media/wwwnclacuk/centreforruraleconomy/files/discussion-pa per-11.pdf

6. Maxwell, S., & Smith, M. (1992). Household food security; a conceptual review. In S. Maxwell & T.R. Frankenberger, Household Food Security: Concepts, Indicators, Measurements: A Technical Review. New York and Rome: UNICEF and IFAD. https://www.drcsc.org/resources/FoodSecurity-Concept%20of%20Food%20Security2 .pdf

7. Robertson, S. (2004), "Understanding inverse document frequency: on theoretical arguments for IDF", Journal of Documentation, Vol. 60 No. 5, pp. 503-520. https://doi.org/10.1108/00220410410560582